IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 June 2022, accepted 15 August 2022, date of publication 26 August 2022, date of current version 2 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3202295

METHODS

Hepatitis C Virus Detection Model by Using Random Forest, Logistic-Regression and ABC Algorithm

TZUU-HSENG S. LI[1], (Member, IEEE), HUAN-JUNG CHIU[1], AND PING-HUAN KUO[2], (Associate Member, IEEE)
[1]aiRobots Laboratory, Department of Electrical Engineering, National Cheng Kung University, Tainan 70101, Taiwan
[2]Department of Mechanical Engineering, National Chung Cheng University, Chiayi 62102, Taiwan

Corresponding authors: Ping-Huan Kuo (phkuo@ccu.edu.tw) and Tzuu-Hseng S. Li (thsli@mail.ncku.edu.tw)

This work was supported by the Ministry of Science and Technology, Taiwan, under Grant MOST 109-2221-E-006-196-MY3 and Grant MOST 109-2221-E-194-053-MY3.

ABSTRACT This study proposes an automatic classifier for detecting the multiclass probabilities of hepatitis C virus (HCV) incidence based on patients' blood attributes. The purpose of this study is to establish an artificial intelligence-based model that can identify HCV patients and detect the disease in early stage for future treatments. This model can be applied by using clinical data and keeps the performance from imbalanced datasets. The innovation in this article lies in considering the ‘‘unbalanced data’’ existing in medical record-based clinical data. Synthetic minority oversampling technique (SMOTE) algorithm was further employed to derive corresponding solutions. This objective was achieved using a cascade two-stage method combining the random forest (RF) and logistic regression (LR) algorithms. Two models were trained by applying the RF (Model 1) and LR (Model 2) to raw and preprocessed data, respectively. The artificial bee colony (ABC) algorithm was then used to determine the optimal threshold value required for filtering and separation, that is, the optimal combination of both models. The two-stage mixing algorithm combines algorithms of different search dimensions, thus integrating the strengths of those algorithms. The critical threshold value for separating Model 1 and Model 2 was obtained through an optimized search using the ABC algorithm. After conducting 10-fold Monte Carlo cross-validation experiments 50 times (for mean values), data from the recent pandemic were used to verify the proposed method. To evaluate the quantitative results, indicators, such as prediction accuracy, precision, recall, F1-score, and Matthews correlation coefficient, were compared with those of the latest algorithms used in relevant fields. The results indicate that the proposed model, named Cascade RF-LR (with SMOTE), can be used to detect the multiclass probabilities of HCV incidence using the ABC algorithm, thereby improving the effectiveness of relevant treatments.

INDEX TERMS Random forest, logistic regression, two-stage mixing, ABC algorithm, 10-fold Monte-Carlo cross-validation, synthetic minority oversampling technique.

I. INTRODUCTION

In medical research, redefining the influence of medical care data can improve medical care quality. In the medical care field, data centers that compile patients' medical records and examination results will serve as crucial factors for improving the quality of medical care for patients [1]. When knowledge is extracted from the mining of medical record data, various perspectives can be adopted to explore disease incidence, progression, and spreading, and such exploration can provide valuable information for ascertaining the diagnosis and treatment of diseases. Therefore, data mining can uncover the underlying relationships, trends, and patterns between data, and in turn enhance the accurate identification of diseases [2].

In this study, the researchers targeted the patients diagnosed with liver diseases and defined hepatitis as the liver

The associate editor coordinating the review of this manuscript and approving it for publication was Huiyan Zhang.

inflammation from any cause resulting in damage of liver cells. When external substances or pathogens invade the human body, the immune system activates inflammatory cells (e.g., lymphocytes), which infiltrate into tissues and release immune substances to fight the invaders. This condition is known as the inflammatory response, or inflammation in laymen's terms. Hepatitis is mainly divided into two types: viral and non-viral. The types of viral hepatitis include hepatitis A, B, C, D, and E.

Chronic hepatitis C virus (HCV) infection is one of the main causes of liver cirrhosis and hepatocellular carcinoma worldwide [3]. It increases the mortality and incidence rate of hepatic and extrahepatic diseases, particularly in patients with HCV viremia. Furthermore, alcoholic liver disease—, which progresses from mild liver disease to alcoholic hepatitis, and finally to cirrhosis—, is the main cause of global hepatitis incidence and mortality [4]. In Taiwan, the prevalence rate of HCV is 2.1%, which translates to a population of 489,000 patients with HCV viremia [5]. Moreover, during viral pandemics, patients with chronic liver diseases pose a huge challenge to the medical health care systems [6]. HCV belongs to the Hepevirus family, and hepatitis C is caused by HCV infection. After an acute infection, approximately 20%–30% of patients would exhibit clinical symptoms such as fever, fatigue, loss of appetite, slight abdominal discomfort, nausea, vomiting, jaundice, and other related symptoms [7]. The severity of HVC-related diseases can range from unobvious symptoms to the deadly fulminant hepatitis. Therefore, preventing the transmission of HCV is crucial, for which blood tests and screening for other variables are highly beneficial.

Currently, numerous liver disease diagnostic methods are based on machine learning. Several methods that have been used to examine pathological changes in hepatitis are described as follows: In [8], computed tomography was used to automatically locate the healthy segment of the liver and the segment with lesions using a modified method called CALOFCM, which combines fast fuzzy C-means (FCM), chaos theory, and the bioinspired ant lion optimizer (ALO). The chaos theory-based ALO prevented FCM from falling into the local minimum, enhanced the calculation performance, improved stability, reduced the sensitivity of the iteration process, and allowed the use of the optimal barycenter through FCM. In [9], ultrasound images of chronic liver diseases, laboratory examination results, and clinical records were used to perform auto classification of chronic liver disease stages. Specifically, a clinical-based classifier was first used to separate healthy conditions from pathological conditions. When an unhealthy condition was detected, this method classified the results into three types of exclusive pathologies: (1) chronic hepatitis, (2) compensated cirrhosis, and (3) decompensated cirrhosis. The features used and classifiers (Bayes, Parzen, support vector machine [SVM], and k-nearest neighbor [KNN]) were optimally selected for each stage [9]. However, there are many powerful optimization algorithms which are used in many research fields, such as dragonfly algorithm [10], ant lion algorithm [10], modified firefly algorithm [11], modified ABC algorithms [11], modified ant colony optimization [12], enhanced firefly algorithm [13], and so on. [13] is applied in the SoC-based test dispatch and time in order to save on the time and cost spent. The enhanced firefly algorithm is used. The performances of these algorithm are validated in the experimental results.

In [14], liver disease datasets were used to evaluate models, data mining models were compared to select critical features for predicting liver diseases, and the extraction, loading, transformation, and analysis method was used to compare different models, namely, random forest (RF), multilayer perceptron (MLP) neural network, Bayesian network, SVM, and particle swarm optimization. In [15], a machine-learning model was constructed based on 2009 clinical data to predict fatty liver disease (FLD). FLD is a clinical complication that commonly occurs during the early phase of chronic liver inflammation (chronic FLD may lead to the chronic inflammation of the liver). The classification models for FLD include RF, naive Bayes (NB), artificial neural networks, and logistic regression (LR). In [16], patients with HCV were analyzed by clinical traits (e.g., age) at first HCV screening, insurance at first HCV screening, race, gender, presence of fibrosis and/or cirrhosis, presence of other liver disease, presence of ascites, transplanted liver, presence of other types of liver cancer, presence of steatosis, presence of liver cell carcinoma, and ethnicity. The three care methods were modeled using decision trees and random forests. The methods were linkage to nursing care, initiation of antiviral treatments, and virologic cure. Furthermore, in response to the worldwide threat posed by COVID-19, clinical studies on the use of machine learning algorithms to combat the spread of the COVID-19 virus have applied virtual filters and machine learning algorithms to identify new drug candidates [17] and conduct the drug repurposing of anti-hepatitis C drug derivatives for COVID-19 treatment [18]. Furthermore, some articles [19], [20] also applied several artificial intelligence techniques for the liver disease detection. In addition, [21] gave attention to recent breast cancer disease topics that used machine learning methods. In addition, in exploring the diagnosis of Alzheimer's disease [22], the volumetric feature-based sMRI data of hippocampal slices was used. The convolutional neural network and deep neural network were adopted. In Article [23], the latest machine learning and deep learning method applied to detect four brain diseases: Alzheimer's Disease (AD), brain tumor, epilepsy and Parkinson's Disease were reviewed. In addition, different machine learning and deep learning methods, models, data sets, etc. were taken into account.

The main contributions of this study are as follows:(1) A two-stage joint model, in which the RF and LR models (Model 1 and Model 2, respectively) were integrated with the artificial bee colony (ABC) algorithm, and the synthetic minority oversampling technique (SMOTE) and feature selection method were also used to improve the model fit for Model 2, is constructed. (2) The proposed method addresses the problem of imbalanced data inevitably appearing in clinical data, which was not considered by other algorithms. (3) A diversified range of indicators is used to evaluate the proposed model under different evaluation needs. (4) Verification based on the mean values obtained through 10-fold Monte Carlo cross-validation experiments performed 50 times is performed and the scores are compared with those obtained using the latest algorithms.

The remainder of this article is as follows: Section II explains the major components of the proposed algorithms. Section III presents the proposed methodology in detail. Section IV provides the experimental results and discussion. Finally, Section V presents the conclusions of the present study.

II. RELATED WORK

This chapter introduces the methods used in the present study and their latest medical and clinical applications. These include the conventional RF, LR, ABC, and SMOTE methods.

A. CASCADE CLASSIFIERS

A cascade classifier is a classification method involving the combination of complicated classifiers and is often used in image object detection. A cascade classifier can rapidly discard the background of images to spend more calculation resources on the more hopeful target region; cascading can be regarded as a target-specific focus mechanism [24]. A cascade classification model is a type of joint classification model that combines a set of the latest classifiers to improve the results they produce, and sharing of information between tasks is achieved through the linkage of component classifiers [25].

A cascade classifier is a great tool for processing extremely imbalanced data (i.e., data with too many negative numbers and too few positive numbers [26]). One of the most recent studies [27] investigated the design of complexity-aware cascade pedestrian detectors.

In the field of biochemistry, cascade classifiers have been used in physical biochemistry networks. In a case study [28], the researchers proposed a cascade learning framework that incorporated semantic features from a knowledge embedding model and graph features from a graph embedding model. This framework combined the features into a single architecture that fully utilized the advantages of the two feature types. The case study empirically demonstrated the value of this framework in identifying potential relationships between diseases, drugs, genetics, and treatment methods. In neurology and clinical studies, cascade classifiers were used in the auto-evaluation of subjects' neurocognitive performance [29], which was achieved through the analysis of electroencephalographic signals. The cascade framework was composed of two long short-term memory recurrent neural networks.

B. RF

The RF method proposed by Breiman in 2001 [30] has achieved great success as a general classification and regression method. This supervised learning procedure operates according to a simple but effective divide-and-conquer principle: First, sampling is performed on data, and a random tree predictor is ''grown'' on each fragment. Then predictions can be made based on the mean values generated by these predictors. The RF method has become popular owing to its applicability in an extensive range of prediction problems. Apart from being simple and easy to use, this method is well known for its accuracy and competency in handling small samples and high-dimensional feature spaces. Moreover, it can easily be used in parallel with other algorithms, endowing it with the potential to realize large-scale reality processing systems [31].

In medical applications, RF is used to extract the important features of electrocardiogram signals for the classification of different arrhythmias [32]. RF is also used in the correct classification of Cushing's syndrome. In particular, it is used in promoting treatments and improving prognosis for patients with Cushing's syndrome. A relevant study indicated that RF is the most suitable method for classifying the syndrome [33]. Regarding the high costs involved in the prediction of treatment fees for patients with asthma, the frequently used comorbidity portfolio design involves the recombination of comorbidities in different budgets, where the training for comorbidity portfolio design includes the training of RF prediction models [34]. To resolve the class-imbalanced data problem in data classification, especially the lack of identification for minority groups, the class-weight RF method was introduced to assign a single weight value for each class [35].

C. MULTICLASS LR

Multiclass LR is an algorithm that is particularly suitable for the discovery of features or the associations between certain specific results: LR is a type of probabilistic classifier, differentiating it from the purely generative classifier (NB) or purely discriminative classifier (LR). In natural language processing, LR is a baseline-supervised machine-learning algorithm used for supervisory purposes, and it is closely related to neural networks.

Neural networks can be regarded as a series of LR classifiers piled on top of a logical network [36]. Assume there are n-th training instances inputting/outputting data to (x_i, y_i); then X $= (x_1, x_i \ldots, x_n) \in R^{d \times n}$, Instance x_i, in which the attributes are d-dimensional vectors. For input x_i, the feature vector $x_i = [x_i^1, x_i^j, \ldots, x_i^d]$, and feature j will be named x_i^j. This classification problem is resolved through the learning of weight vectors and bias terms from the training set. The sigmoid function (softmax function would be used for multiclass) would be used to calculate the probability of $p(y|x)$, which is then used to estimate the category of y.

$$\hat{y} = sigmoid(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

$$softmax(z) = [\frac{e^{z_1}}{\sum_{i=1}^{k} e^{z_i}}, \frac{e^{z_2}}{\sum_{i=1}^{k} e^{z_i}}, ..\frac{e^{z_k}}{\sum_{i=1}^{k} e^{z_i}}] \quad (2)$$

$$z = (\sum_{i}^{n} w_i x_i) + b \quad (3)$$

The classifier would multiply each x_i with weight value w_i, sum up the weight feature, and add error term b, thus obtaining the expressed weighted sum of the category z. The purpose of setting a learning target is to minimize the error in the training samples to the greatest extent, and the formula used is the cross-entropy loss function equation. The result is the cross-entropy loss L_{CE}, which is expressed in Eq. (3):

$$L_{CE}(\hat{y}, y) = -[y \log \hat{y} + (1 - y) \log(1 - \hat{y})] \quad (4)$$

After simplifying $\hat{y} = sigmoid((\sum_{i}^{n} w_i x_i) + b)$, it is substituted into $\sigma(w^T x + b)$:

$$\begin{aligned} L_{CE}(w, b) = &-[y \log \sigma(w^T x + b) \\ &+ (1 - y) \log(1 - \sigma(w^T x + b))] \end{aligned} \quad (5)$$

D. ABC ALGORITHM

The ABC algorithm is based on swarm intelligence and is often used to solve optimization problems; this method is inspired by the food foraging behaviors of bees [37]. Specifically, searching for food sources and locating food indicate possible solutions. The searching mechanism of this method involves three types of bees, namely employed bees, onlooker bees, and scouting bees. They work together so that the location of food sources can be determined in the iteration process of ABC. The employed bees and onlooker bees each constitute half of the population, and their roles are interconvertible. Onlooker bees represent the greed mechanism of ABC; these food foragers play different roles in the ABC algorithm [38].

Based on the estimated probability of the food sources, they will be appointed as the food and source locations, and work involving the development of food sources will be allocated according to these locations. Once an onlooker bee is appointed as the food source, it is converted into an employed bee. The employed bees represent the developing part of the ABC algorithm; they perform searching around the target food sources. Scouting bees are only sent out when a food source has been used for a continuous period because of the lack of better food sources. Scouting bees represent the searching mechanism of the ABC algorithm. Sending scouting bees to explore brand new food sources ensures that the ABC algorithm can break out from the local optimum.

E. SMOTE

SMOTE is an oversampling technique [39] used for resolving problems caused by imbalanced data, and it is often included as part of machine learning. SMOTE can freely create new minority class examples from the nearest neighbors of minority-class samples. The new instances are created by inserting new instances in the KNNs, and this process would not affect the distribution of the original source data.

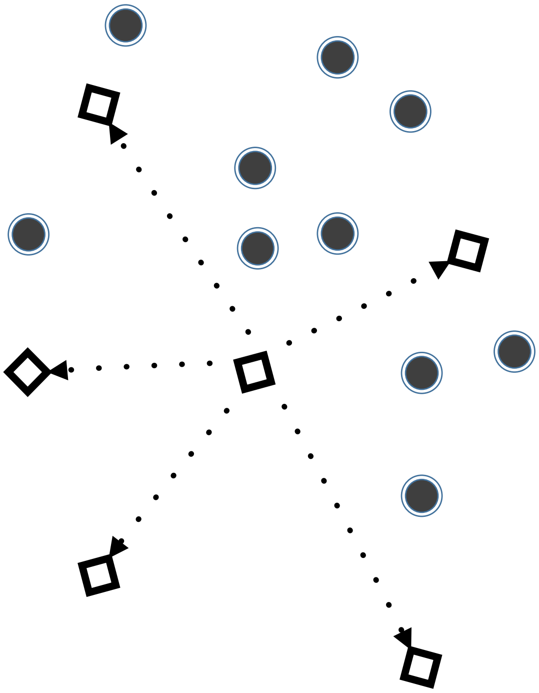


FIGURE 1. **Creation of the minority class synthetic data point using SMOTE. Sampling points were introduced by obtaining the line segments of the k-th nearest neighbors of each minority sample; the neighbors were randomly chosen from the neighborhood of the k-th nearest neighbors according to the number of samples required. Square represents the selected point for the minority class, and dotted lines represent the possible synthesized data points constructed through random interpolation.**

$$x_{new} = x_{select} + (x_{nearest} - x_{select}) \times \delta; \quad \delta \in [0, 1] \quad (6)$$

This method can be used to eliminate the harmful effect of a skewed distribution [40]. These new examples are created based on the features of the original dataset; the purpose of creating them based on the features of the original dataset is that they will be similar to the original minority examples, which also prevents the occurrence of sampling bias [41], [42]. This method has been used in synthesizing minority samples in the medical field [43]; a further explanation of this technique is presented in Fig. 1 and concept presented in Eq. (6).

F. FEATURE SELECTION

In many classification tasks, feature selection is a crucial method in reducing the dimensionality of data in the preprocessing phase because these irrelevant and excess features would mislead the learning process; this is dependent on the chosen method.

III. METHOD

This section mainly explains the two type Models was generated and most crucially the method proposed in this study, namely Cascade RF-LR (with SMOTE) using the ABC algorithm. The cascade two-stage method uses the ABC algorithm to search for the optimal threshold value to connect the two models. Two models were trained by applying the RF (Model 1) and LR (Model 2) to raw and preprocessed data, respectively. The ABC algorithm was then used to determine the optimal combination of both models. Model 1 was an RF Model trained with raw data, whereas Model 2 involved processing the raw data with feature selection and SMOTE to form the training data for the LR model.

A. DATA USED TO CREATE MODEL 1 BY RF ALGORITHM

The purpose of training the RF Model with raw data is to obtain the estimated confidence probability of entities within the verification data and the weights of data features. Decision tree nodes were chosen at random to divide the features; consequently, model training was efficient when the sample features were highly dimensional. After training, the model could also yield the importance of each feature to the output; this information was used again in the training data of the LR model. In Eq. (7), Model 1 is expressed as $model_{RF}(x_n^{original}) = y_n^{original}$, and the corresponding training data is expressed in Eq. (8) as $D = \{X^{original}, y^{original}\}$; the corresponding confidence probability estimates derived using the established RF Model is expressed as $model_{RF}^{probabilities}(D_{val})$, with D_{val} representing the input verification dataset in Eq. (10).

B. DATA USED TO CREATE MODEL 2 BY MLR ALGORITHM

Although the RF Model already provides a certain level of performance, data imbalances will inevitably appear in the clinical data of medical cases. The RF Model did not perform optimally in the subsequent performance evaluation experiments and was prone to overfitting when it was processing specific samples with high noise levels. Consequently, the raw data was preprocessed to differentiate the training data for Model 1 and to identify the corresponding relationships between features; preprocessing also solved the problem of data imbalance. Therefore, data preprocessing included the application of feature selection and SMOTE to the raw data.

1) FEATURE SELECTION BY THE RF MODEL

When the bagging method was applied on the component classifier algorithm during RF model training, different training datasets were generated using bootstrap sampling for the purpose of constructing different classifiers. These data are known as the out-of-bag (OOB) data. The OOB data were used to calculate the importance of each feature. After the data were subjected to the feature selection process, they were further passed to the LR model for model construction.

2) SOLVING LR-MODEL SKEWNESS DISTRIBUTION USING SMOTE

Minority-class data being used for second-stage model training would have resulted in a prominent impairment in accuracy. To eliminate the harmful effects of skewed distribution, the over-resampling technique was used to fill up the data for the minority class. SMOTE is one of the most renowned techniques for resolving this problem in the field because it enables the establishment of a model under the condition of balanced data.

After the raw data were preprocessed through the aforementioned steps, the LR model (i.e., Model 2) was built using the samples and the LR method. In contrast to the random sampling that is conducted during the application of the bagging method in the RL Model, LR is an algorithm that is particularly suitable for identifying the features of or the associations between specific results. LR is a type of probabilistic classifier, and it is one of the most widely applied machine learning algorithms. Logistic regression is the most straightforward algorithm to understand and apply to combinations of two different types of models, and its computational cost is low. Model 2 is expressed in Eq. (7) as $model_{LR}(x_m^{preprocess}) = y_m^{preprocess}$; the corresponding training data were expressed in Eq. (8) as $D^{preprocess} = f_{SMOTE}(f_{feature\_sele}(D))$.

$$\begin{aligned} model_{RF}(x_n^{original}) &= y_n^{original}, model_{LR}(x_m^{preprocess}) \\ &= y_m^{preprocess} \end{aligned} \tag{7}$$

$$\begin{aligned} D &= \{X^{original}, y^{original}\}, D^{preprocess} \\ &= f_{SMOTE}(f_{feature\_sele}(D)) \end{aligned} \tag{8}$$

$$\begin{aligned} f_{i*}^{ABC} &= \underset{L \leq i \leq U}{argmax}[model_{RF}(D_{val} - D_{val,i*}^{-}) \\ &\quad + model_{LR}(D_{val,i*}^{-})] \end{aligned} \tag{9}$$

$$D_{val,i*}^{-} = model_{RF}^{probabilities}(D_{val}) \leq i \tag{10}$$

C. CASCADE RF–MLR BY THE ABC ALGORITHM

The cascade two-stage mode identifies the optimal threshold value using the ABC algorithm, such that this value can act as the linkage between models, as shown in Fig. 2. First, two models were trained using the RF and LR methods and the original and preprocessed training data (in Algorithm 1 line 3), and the optimal combination for the two models was identified using the ABC algorithm (in Algorithm 1 line 15). Imbalanced data inevitably appear in the clinical data collected from medical cases. The method for using this combination was not considered in the case of other combinations.

The data preprocessed using feature selection and the SMOTE method (in Algorithm 1 line 5). The estimated confidence probability was first identified using the verification data and RF model. It is expressed in Eq. (10) as $model_{RF}^{probabilities}(D_{val}) \leq i$, where D_{val} represents the verification data, and i is the estimated value of the optimal confidence probability selected by the ABC algorithm. Then, the threshold value with the optimal probability was selected as the basis for data separation, and the separated data were passed to the LR model for further judgment. Lastly, the ABC algorithm was used to identify the optimal threshold value, which was used in identifying the optimal predicted classification for the cascade two-stage model. The ABC algorithm

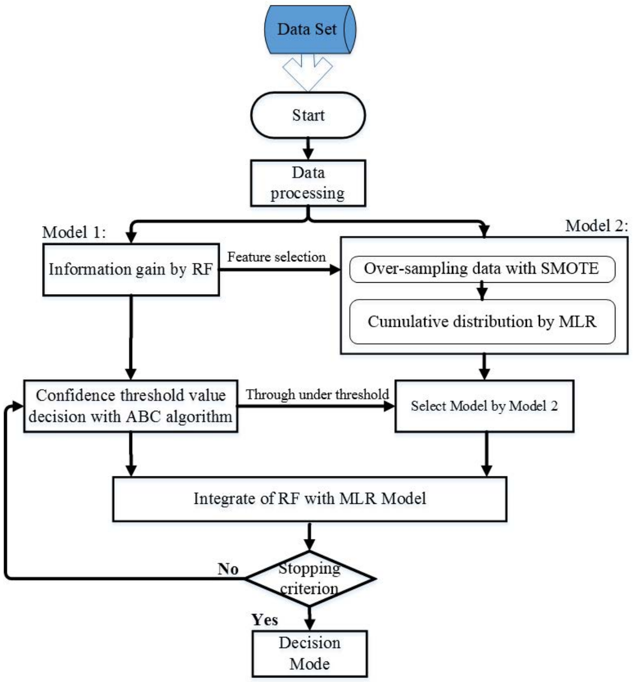


FIGURE 2. System flowchart of Cascade RF–LR (with SMOTE) using ABC algorithm.

TABLE 1. Description of similar IR datasets.

Data Set	C.	Attr.	Distribution	Size	IR
HCV	4	14	526/20/12/24	582	43.83
Thyroid	3	21	166/368/6,666	7,200	40.16
Page-blocks	5	10	4,913/329/28/88/115	5,473	175.46

object functions are presented in Eq. (9). Here, $D^{-}_{val,i}$ represents the data samples in $model_{RF}^{probabilities}$ screened to be lower the estimated value of optimal confidence probability i and to determine the majority combinations in both models, respectively. Furthermore, the value of i was between L and U and expressed as $L \leq i \leq U$ (in Algorithm 1 line 11); this was obtained directly through the tests performed during the experiment. The RF model and optimal separation threshold value i^{*} were obtained at this stage. As shown in Fig. 2, after combining these two models, the iteration process will be completed when 95% training accuracy is reached.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

This section describes the multi-classification database used for validation and the experiment setup; it also reviews performance measurement and compares the multiclass indicators used in the present study with the latest algorithms used in other relevant studies. These algorithms include RF, deep forest (gcForest) [44], [45], extreme Gradient boosting (XGBoost) [46], decision tree (DecisionTree), KNN, Gaussian NB (GaussianNB), and partial least squares two-block regression (PLS2Regression) [47], [48].

Algorithm 1 Pseudocode of the Proposed Cascade RF-LR Process [25]

```
1:  $D = \{X^{original}, y^{original}\}$ ←instance of raw data using Eq. (8).
2:  function genTwoModels (Argument$D$){
3:  //Two models were trained by applying the RF (Model 1) and LR (Model 2) to raw and preprocessed data, respectively
4:      $model_{RF}(x_n^{original}) = y_n^{original}$ ← Model 1 was an RF algorithm trained with raw data using Eq. (7).
5:      $D^{preprocess} = f_{SMOTE}(f_{feature\_sele}(D))$ ← processing the raw data with feature selection and the synthetic minority oversampling technique (SMOTE) to form the training data for the LR model using Eq. (8).
6:      $model_{LR}(x_m^{preprocess}) = y_m^{preprocess}$ ← Model 2 was trained by applying the LR algorithm to preprocessed data using Eq. (7).
7:      return RF (Model 1), LR (Model 2)
8:  }
9:  // Confidence threshold value decision with ABC algorithm
    Initialization: // Artificial Bee Colony (ABC) Algorithm
10: $L, U$ ← Set search solution range between $L$ and $U$
11: $f_{i*}^{ABC} = \underset{L\leq i\leq U}{argmax}[model_{RF}(D_{val} - D^{-}_{val,i*}) + model_{LR}(D^{-}_{val,i*})]$ ← Set object functions using Eq. (9).
12: Set maximum number of iteration =50,
13: Set the population size = 50; // where population size = onlookerBee = mpolyeedBee;
14: function ABC-Algorithm (objectFunctions, searchRange, populationSize, maxIteration){
15:     // Integrate of RF with MLR Model
16:     return bestSolution = $i$; //ABC algorithm was used to identify the optimal threshold value
17: }
18: End // Detection Model is when best solution = $i$, The final results were a combination of the results obtained by the two models
```

A. DATABASE

The HCV data used in this study were taken from the Machine Learning Repository of the University of California, Irvine (UCI) [49], [50], [51]. The dataset originally contained a total of 615 instances, four classes, and 14 attributes.

The elimination of some missing values resulted in 582 remaining instances. This dataset has clear data-imbalance problem; specifically, a great discrepancy exists between the sample size of the class with the highest and lowest sample number, making this dataset an imbalanced dataset with an imbalance ratio (IR) of 43.83. Two additional datasets were used as comparison datasets to verify the proposed method: data sets containing multiple classes and

TABLE 2. Corresponding parameters of the algorithms.

Algorithm	Settings	Parameters
RF	*n_estimators*	100
	criterion	gini
	max_depth	None
	min_samples_split	2
	max_features	auto
	bootstrap	True
	oob_score	True
SMOTE	*k_neighbors*	4
LR	*penalty*	L2
	tol	0.0001
	C	1.0
	fit_intercept	True
	max_iter	7000
ABC	*num_employers*	50
	L, U	0.7, 0.9

minority classes were specially selected for this purpose. The IR value and other attributes of the datasets are presented in Table 1, and these data were also retrieved from the Machine Learning Repository of UCI.

B. EXPERIMENT SETUP

The selected verification dataset was first checked for samples with missing features, which were then removed. Next, k-fold cross validation was performed. Specifically, the data were randomly divided into k sets, of which one was selected to be the testing data, and the others were designated as training data. These steps were repeated until each set had been designated as the testing data, that is, k tests had been performed. If we set k to be 10 in the experiment, then a 10-fold cross-validation was performed. This validation was run 50 times to determine the averages and to verify the robustness of the proposed method. In addition, during the training process, one-tenth of the training datasets were used as validation data, which were used to evaluate the performance of the overall validation method. Table 2 presents the algorithms and parameter settings.

C. PERFORMANCE MEASUREMENT

In machine learning, a task that involves two or more classification tasks is known as a ‘‘multiclass classification’’ task. The dataset used in this study was based on multiclass classification, and the problem of minority groups was taken into consideration. This section presents the performance measurement standards selected by the researchers, which were used to assess the proposed multiclass classifier. The selected measurement methods, which are presented in Table 3, were as follows: accuracy, precision, recall, F1-score and Matthews correlation coefficient (MCC).

Accuracy refers to the probability of the model making a correct prediction. In the case of whole-sample prediction, accuracy refers to the measurement of the model making correct predictions for all classes. The precision and recall indexes for each category would need to be calculated separately. Precision is the measurement of accuracy; in other words, it is an indicator of how many samples labeled as positive are correctly labeled [52] (i.e., positive predictive values). When the cost of false positives is high and their occurrence is expected to be minimized, the enhancement of precision measurement values should be emphasized. Furthermore, this indicator can reflect the precision level of each response class, and therefore, it is suitable to be used on minority classes [53].

Recall is an indicator of how many positive samples are correctly labeled. When the cost of false negatives is high and their occurrence is expected to be minimized, the enhancement of recall measurement values should be emphasized. In multiclass classification, recall represents the percentage of positives in class K that are correctly identified (i.e., the true positive rate). In other words, recall is an indicator for measuring integrity (as plotted in Fig. 3). The F-score is a measurement method that combines the two indicators of precision and recall and is used to express the weighted mean of the two indicators. In cases of uneven class distribution, the F-score is often more useful than accuracy.

MCC is a correlation coefficient [54] indicating the correlation between the observed and predicted classifications, it returns a value that ranges between -1 and $+1$. A coefficient of $+1$ represents a perfect prediction, a coefficient of 0 represents a prediction that is no better than random; and a coefficient of -1 represents a total disagreement between prediction and observation results. Furthermore, MCC can be used when the size difference between categories is huge. In recent years, it has become an extensively utilized measurement standard in the testing of machine-learning performance [55], and it is suitable for use in targeting different classes under a multiclass condition [56].

1) ACCURACY SCORE

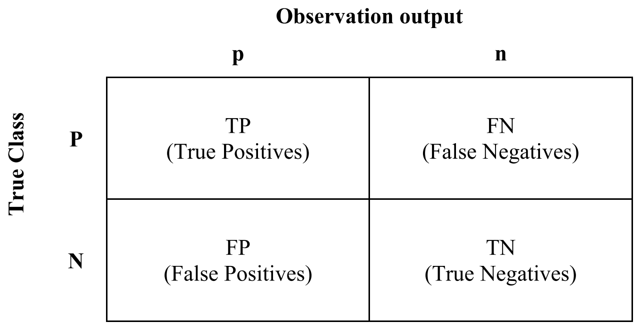
As shown in Fig. 4, the model established based on the accuracy indicator only took into consideration the ratio of the correctly identified instances among the number of classes. This indicator did not consider the class differences. A discrepancy of 2.04% was observed between the average performance of the proposed method and the second-best algorithm (i.e., XGBoost), and a discrepancy of 1.38% was observed between the best performance of the proposed method and the second-best algorithm. This result indicated that the proposed method exhibited greater performance than the other methods in terms of accuracy scores.

2) PRECISION SCORE

The weighted-average precision (as plotted in Fig. 5) measures the weighted mean of each category. In consideration of the class differences, the weighted mean value of each class indicator was calculated separately based on the weight of

TABLE 3. Functions of the classification performance measures.

Symbol	Metric	Defined
ACC	Accuracy	$\frac{TP+TN}{(TP+TN+FP+FN)}$
$Precision_k$	Precision,	$\frac{TP_k}{(TP_k+FP_k)}$
$Recall_k$	Recall, Sensitivity	$\frac{TP_k}{(TP_k+FN_k)}$
$F1\text{-}measure$	F1-score	$\frac{2\times Precision\times Recall}{Precision+Recall}$
MCC	Matthews Correlation Coefficient	$\frac{c\times s-\sum_k^K p_k\times t_k}{\sqrt{(s^2-\sum_k^K p_k^2)(s^2-\sum_k^K t_k^2)}}$



Observation output

True Class		p	n
	P	TP (True Positives)	FN (False Negatives)
	N	FP (False Positives)	TN (True Negatives)

FIGURE 3. Performance evaluation using a confusion matrix.

each class (i.e., the total sample size of each class), as shown in Eq. (11). In this regard, the performance of some algorithms would be slightly superior to the accuracy score after the multiclass factor is considered. If the problem of data balance is not considered, then the unweighted mean values would be used (as plotted in Fig. 6). The results revealed that only three algorithms had a vertical axis legend position higher than 0.8. Moreover, when the number of classes was low, the overall accuracy dropped, and the average performance of Cascade RF–LR (with SMOTE) using the ABC algorithm was only 77.11%. For further discussion on the numerical values of other classes, please refer to Table 4. Results presented in this table indicate that in class 1, the proposed method exhibited greater performance compared with the other algorithms, whereas SVC and gcForest exhibited the greatest performance in class 3.

3) RECALL SCORE

In multiclass classification, recall represents the ratio of correctly identifying a class among all classes. Similarly, weighted-average recall (as plotted in Fig.7) is the measured weighted mean. Because this is a weighted mean value, large quantity differences between class weights would exist in class performance. According to the derivation based on Eqn. (13), the performance of recall score is consistent with the performance of accuracy score. Considering the unbalanced data of each class, the unweighted mean values were used (as plotted in Fig. 8); the maximum indicator for the vertical axis legend position dropped from 1.0 to 0.75. The results indicated that the average performance of the proposed method (i.e., Cascade RF-LR (with SMOTE) using the ABC algorithm) was 71.53%, which exceeded the average indicator of the second-best algorithm by 9.2%. For further discussion on the numerical values of other classes, please refer to Table 5.

4) F1 SCORE

The F1-score is a measurement approach that combines two measurement methods, namely precision and recall. In essence, F1-score is the harmonic mean of precision and recall. They are expressed in weighted mean values (Fig. 9) and unweighted mean values (Fig. 10). As depicted in both figures, the performance of this algorithm surpassed that of the other algorithms. For further discussion on the numerical values of other classes, please refer to Table 6.

5) MCC SCORE

In essence, MCC is a correlation coefficient that ranges between −1 and +1; a correlation coefficient of +1 indicates a perfect prediction, whereas −1 indicates an inverse prediction. As depicted in Fig. 11, the average performance of Cascade RF–LR (with SMOTE) using the ABC algorithm was 78.84%, which surpassed the performance of the second-best algorithm by 10.33%.

D. K-FOLD-MONTE CROSS-VALIDATION

We performed 50 runs of 10-fold Monte Carlo cross-validation to compare the indicators and other algorithms, 9/10 of all the data were used for training purposes, whereas the remaining data were used for verifying the chosen methods. Moreover, 1/10 of the data were randomly retained every turn to avoid overfitting and selection bias. The k-fold values for each turn were saved, and the final data after 50 runs are illustrated in the figures below. In addition, the final mean and standard deviation values of the 50 runs are presented.

E. COMPARISON WITH OTHER DATASETS

For the performance of the other two datasets, please refer to Table 7 and Table 8. The results of all indicators indicated that the proposed method exhibited the optimal performance among the tested methods. The only exception was found in the results of macro-average precision. In the thyroid dataset, a discrepancy of 5.06% existed between the proposed method and the best-performing indicator. In the page block dataset, a discrepancy of 20.53% existed between the proposed method and the best-performing indicator. In particular, the same algorithm was not used in the indicators with greater performance. However, among the weighted-average precision indicators, the proposed method exhibited the optimal performance.

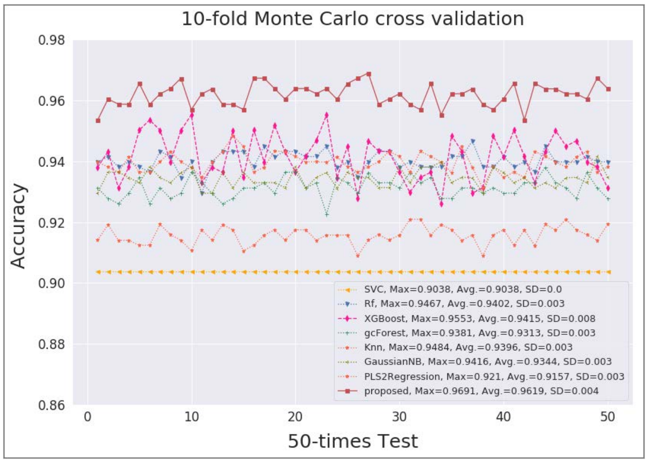


FIGURE 4. Average accuracy comparison between the proposed method and other state-of-the-art methods with 50 runs of 10-fold Monte Carlo cross-validation.

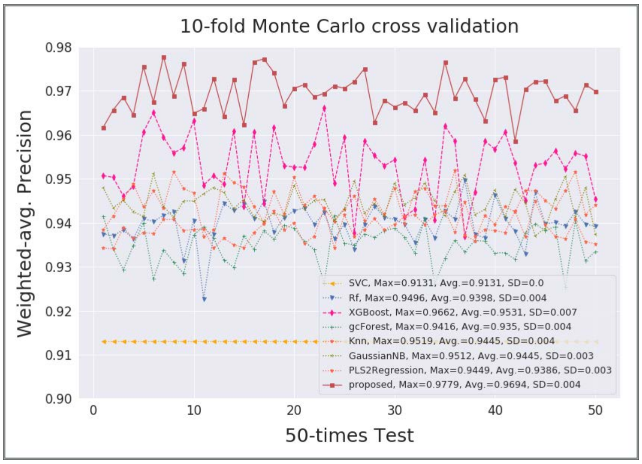


FIGURE 5. Weighted-average precision, which individually calculated each class and estimated a weighted mean of the measures; the proposed method was compared with other state-of-the-art methods through 50 runs of 10-fold Monte Carlo cross-validation.

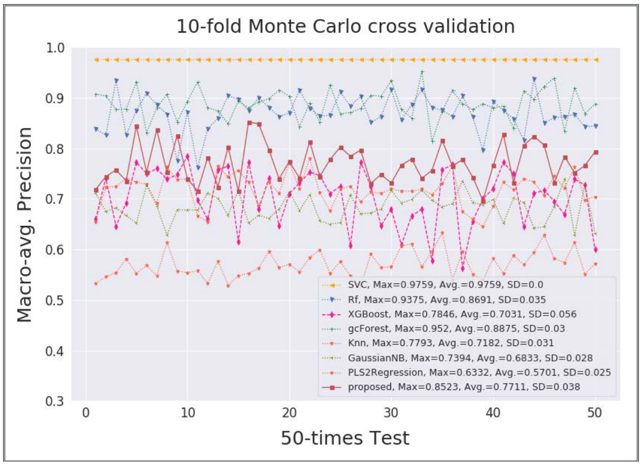


FIGURE 6. Unweighted-average precision, which individually calculated each class and estimated an unweighted mean of the measures; the proposed method was compared with other state-of-the-art methods through 50 runs of 10-fold Monte Carlo cross-validation.

F. DISCUSSION AND ANALYSIS

The comparison of the proposed method and the original RF algorithm is depicted in Fig. 4. With regard to the accuracy indicator, the constructed models only consider the ratio of entities correctly identified by category to the overall number of specimens; however, this indicator does not consider differences in category. The average performance results indicate that the proposed method outperformed the original RF algorithm by 2.17%. The above analysis targeted the average accuracy indicator in Fig. 4 where the performance was superior to that of the original RF algorithm, as well as other algorithms compared.

Precision is an indicator of accuracy, and it refers to the number of samples that were correctly marked as positive. The weighted averages (Fig. 5) indicate that the proposed method outperformed the original RF algorithm by 2.96% on average. However, if the data-imbalance problem was not considered, the unweighted averages were used (Fig. 6), and the sum of all the category values was divided by the number of categories, then the original RF algorithm outperformed the proposed method by 9.8% in terms of average performance. In order to analyze different types of and individual unbalanced problems and the differences arising from the indicators in Fig. 5 and Fig. 6, Table 4 was generated and analyzed.

Table 4 was created to analyze this phenomenon and to further discuss the effects of categories with fewer samples. This table compares the accuracy of each category. Compared with Table 1 where the numbers of samples in each category was compared, Table 4 indicates that Class 1 had the greatest number of HCV datasets, and that Class 3 had the fewest; the IR was 43.8. A comparison of the algorithms revealed that the proposed method performed the best in Class 1 but the second worst in Class 3. However, a comparison with the recall indicators (Table 5) indicated that the proposed method had the best performance. The performance in Class 2 and 3 with the least quantities was superior to that of other algorithms compared.

$$weighted\_averaged = \frac{\sum_{k=1}^{n}(|y_k| \times \frac{TP_k}{TP_k+FP_k})}{\sum_{k=1}^{n}|y_k|} \tag{11}$$

$$Marco\_averaged = \frac{1}{n}\sum_{k=1}^{n}(\frac{TP_k}{TP_k+FP_k}) \tag{12}$$

$$\begin{aligned} weighted\_averaged &= \frac{\sum_{k=1}^{n}(|y_k| \times \frac{TP_k}{TP_k+FP_k})}{\sum_{k=1}^{n}|y_k|} \\ &= \frac{\sum_{k=1}^{n}((TP_k+FP_k) \times \frac{TP_k}{TP_k+FP_k})}{\sum_{k=1}^{n}|y_k|} \end{aligned}$$

TABLE 4. Precision average for each class determined through 50 runs of 10-fold monte carlo cross-validation.

Method	Class 1	Class 2	Class 3	Class 4
SVC	0.9038	**1.0**	**1.0**	**1.0**
Rf	0.9501	0.7060	0.8840	0.9362
XGBoost	0.9858	0.5044	0.5096	0.8125
gcForest	0.9445	0.8284	**1.0**	0.7769
Knn	0.9743	0.4235	0.5647	0.9102
GaussianNB	0.9790	0.3604	0.5168	0.8771
PLS2Regression	0.8527	0.3215	0.2776	0.9080
proposed	**0.9930**	0.6743	0.4963	0.9205

TABLE 5. Recall average for each class determined through 10-fold monte carlo cross-validation.

Method	Class 1	Class 2	Class 3	Class 4
SVC	**1.0**	0	0	0
Rf	0.9990	0.1540	0.022	0.7723
XGBoost	0.9836	0.4670	0.290	0.7440
gcForest	0.9940	0.079	0	0.7340
Knn	0.9962	0.4180	0.018	0.5990
GaussianNB	0.9754	0.2250	0.4080	**0.8850**
PLS2Regression	0.8611	0.3867	0.1110	0.2970
proposed	0.9952	**0.4720**	**0.5430**	0.8510

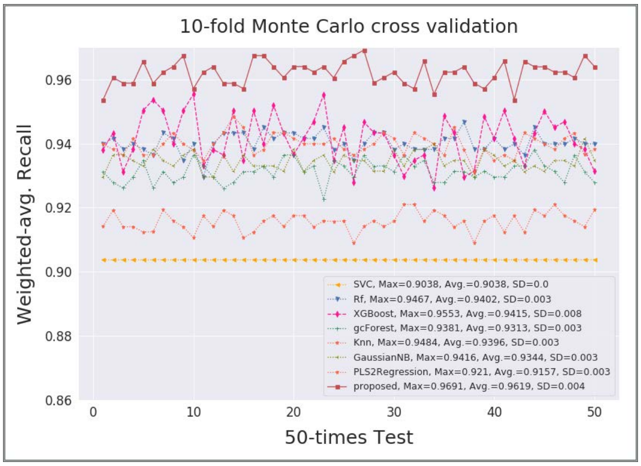


FIGURE 7. Weighted-average recall comparison between the proposed method and other state-of-the-art methods through 50 runs of 10-fold Monte Carlo cross-validation.

$$= \frac{\sum_{k=1}^{n}(TP_k)}{\sum_{k=1}^{n}|y_k|} = \frac{TP_1 + TP_k + ..TP_n}{y_1 + y_k + ...y_n} = \frac{TP}{(TP + TN + FP + FN)} = ACC \tag{13}$$

This demonstrates that in situations involving insufficient sampling, the proposed method has an accuracy rate of approximately 50% and a recall rate that is slightly greater than 50%. By contrast, the other methods have accuracy rates that are greater than 50% and recall rates that are mostly less than 30%. Table 6 (the harmonic means of the accuracy and

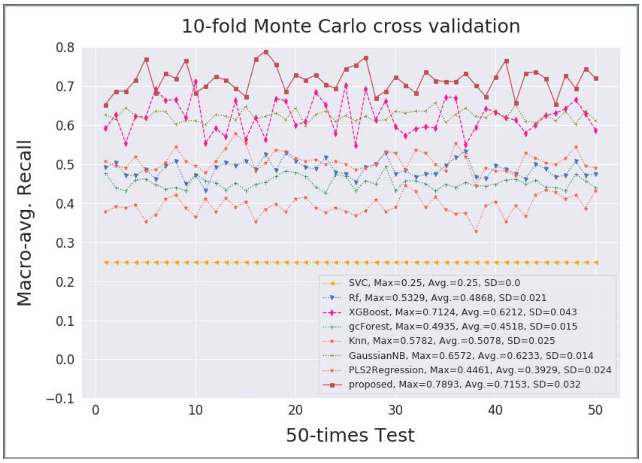


FIGURE 8. Unweighted-average recall comparison between the proposed method and other state-of-the-art methods through 50 runs of 10-fold Monte Carlo cross-validation.

TABLE 6. F1-score average for each class determined through 50 runs of 10-fold monte carlo cross-validation.

Method	Class 1	Class 2	Class 3	Class 4
SVC	0.9494	0	0	0
Rf	0.9738	0.1783	0.022	0.8081
XGBoost	0.9846	0.4175	0.2371	0.7259
gcForest	0.9685	0.096	0	0.7145
Knn	0.9850	0.3606	0.015	0.6667
GaussianNB	0.9770	0.2010	0.3281	**0.8594**
PLS2Regression	0.8568	0.3336	0.087	0.3732
proposed	**0.9940**	**0.4761**	**0.4162**	0.8579

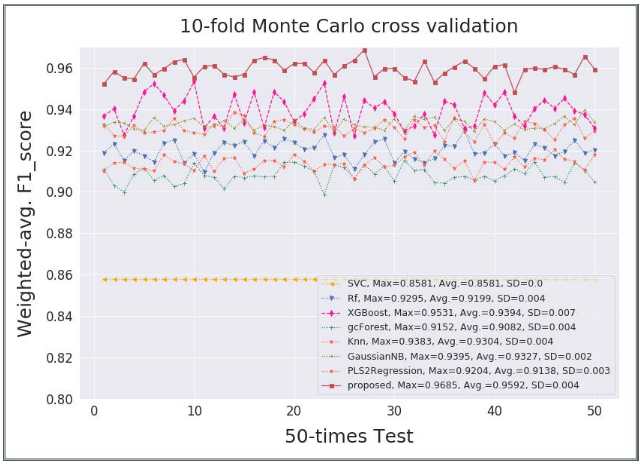


FIGURE 9. Weighted-average F1-score comparison between the proposed method and other state-of-the-art methods through 50 runs of 10-fold Monte Carlo cross-validation.

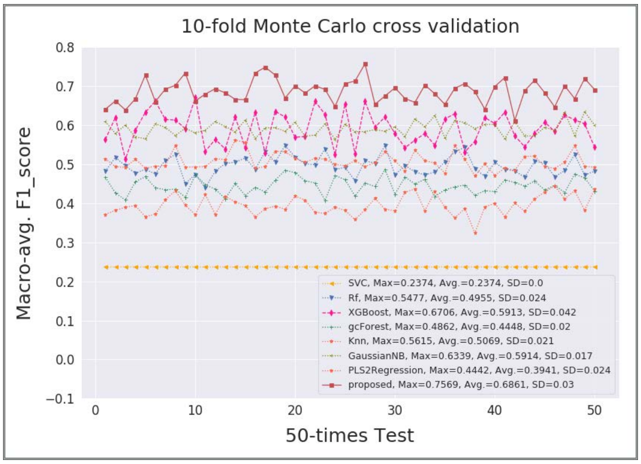


FIGURE 10. Unweighted-average F1-score comparison between the proposed method and other state-of-the-art methods through 50 runs of 10-fold Monte Carlo cross-validation.

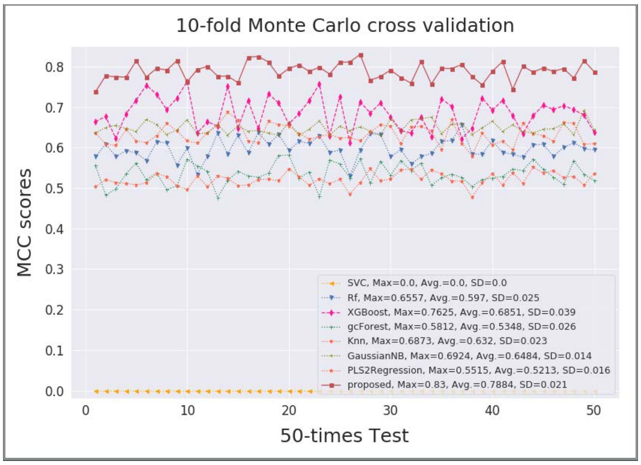


FIGURE 11. MCC comparison between the proposed method and other state-of-the-art methods through 50 runs of 10-fold Monte Carlo cross-validation.

recall rates) reveals that the proposed method produced the best F1 performance in Class 3. The proposed method tends to be conservative in situations involving undersampling, but it does not perform poorly in such situations.

G. ABLATION EXPERIMENTS

In addition, the proposed model consists of several submethods. The ablation experiments could explain the significance of the submethods. In the ablation experiments, the performances of four different submethods combinations (''LR'', ''LR with feature selection'', ''LR with feature selection and SMOTE'', and ''RF'') are validated by 10-fold Monte Carlo cross-validation experiments for 50 times. The final results of ablation experiments in terms of accuracy, weighted-average F1, unweighted-average F1, and MCC-score are illustrated in Fig. 12-15.

Accuracy of the multi-category classifier is not the only indicator for performance evaluation. Thereby, F1-score, which is a combination of precision and recall indicators, is adopted in the experiments. It is noteworthy that the difference between the performances of ''LR with feature selection'' and ''LR with feature selection and SMOTE'' are quite small in Fig. 13. However, the difference between these two above-mentioned submethods is increased by 5% in Fig. 14. This result also indicates that feature selection and SMOTE

TABLE 7. Thyroid dataset average score determined through 10-fold monte carlo cross-validation containing various measurements and comparison of the latest algorithms.

Thyroid								
Method	***Acc.***	$P_{w\text{-}avg}$	P_{avg}	$R_{w\text{-}avg}$	R_{avg}	$F1_{w\text{-}avg}$	$F1_{avg}$	***MCC***
SVC	92.64	93.26	**97.58**	92.74	35.54	89.36	36.04	0.12
Rf	94.33	94.62	96.24	94.33	60.00	91.87	61.36	0.48
XGBoost	93.17	95.21	80.62	93.17	71.72	92.82	69.65	0.60
gcForest	92.58	93.12	97.46	92.58	33.33	89.02	32.05	0.0
Knn	90.61	91.75	64.93	90.61	65.80	91.13	64.96	0.39
Gaussian NB	12.18	92.68	39.57	12.18	61.82	11.51	15.15	0.12
PLS2 Regression	92.55	90.10	44.90	92.55	25.06	90.34	26.56	0.25
proposed	**98.84**	**98.91**	92.50	**98.84**	**95.81**	**98.86**	**93.95**	**0.92**

TABLE 8. Page- blocks dataset average score determined through 10-fold monte carlo cross-validation containing various measurements and comparison of the latest algorithms.

Page- blocks								
Method	***Acc.***	$P_{w\text{-}avg}$	P_{avg}	$R_{w\text{-}avg}$	R_{avg}	$F1_{w\text{-}avg}$	$F1_{avg}$	***MCC***
SVC	91.25	91.02	92.53	91.25	32.88	88.91	36.60	0.37
Rf	95.41	95.42	**95.05**	95.41	50.85	94.14	53.63	0.73
XGBoost	87.45	90.57	69.34	87.45	34.78	85.89	33.32	0.36
gcForest	94.43	94.56	83.85	94.43	56.24	93.60	54.20	0.70
Knn	95.31	94.95	83.99	95.31	62.55	94.73	68.56	0.73
Gaussian NB	88.76	93.02	58.66	88.70	67.68	90.21	57.12	0.52
PLS2 Regression	89.93	91.37	51.67	89.93	31.91	89.26	29.49	0.48
proposed	**96.38**	**96.81**	74.52	**96.38**	**83.14**	**96.50**	**77.28**	**0.82**

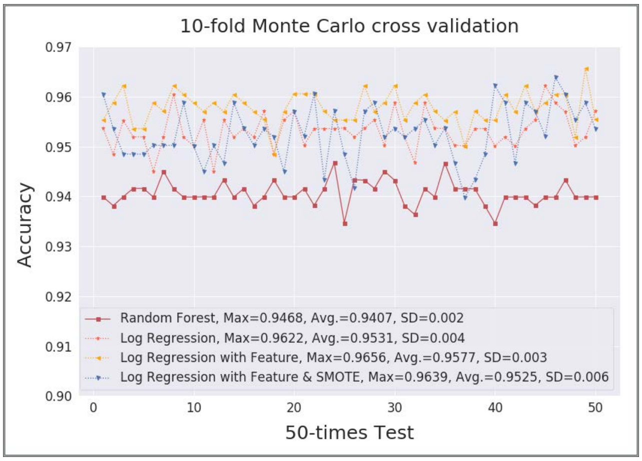


FIGURE 12. Average accuracy comparison ablation experiments with 50 runs of 10-fold Monte Carlo cross-validation.

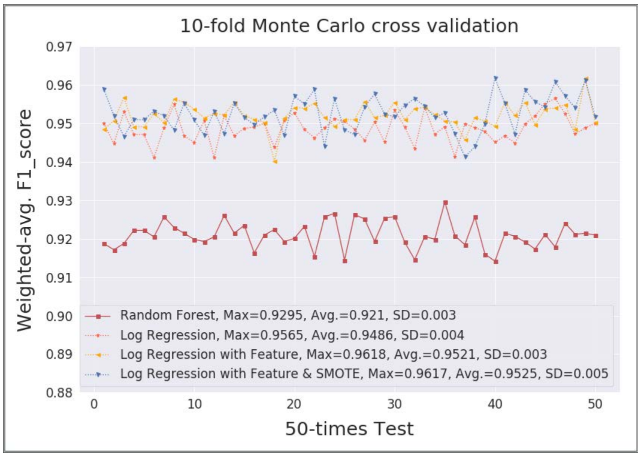


FIGURE 13. Weighted-average F1-score comparison ablation experiments with 50 runs of 10-fold Monte Carlo cross-validation.

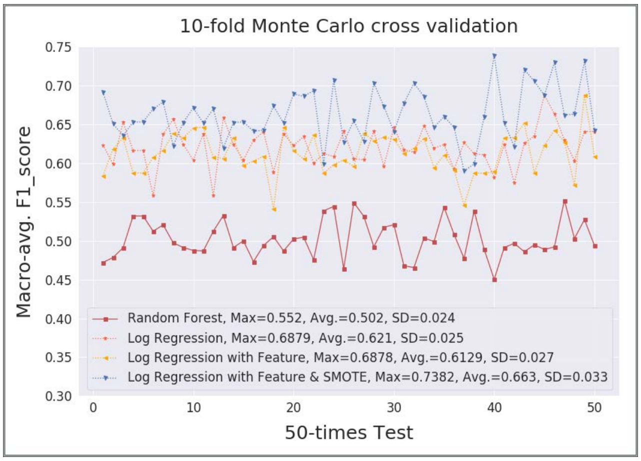


FIGURE 14. Unweighted-average F1-score accuracy comparison ablation experiments with 50 runs of 10-fold Monte Carlo cross-validation.

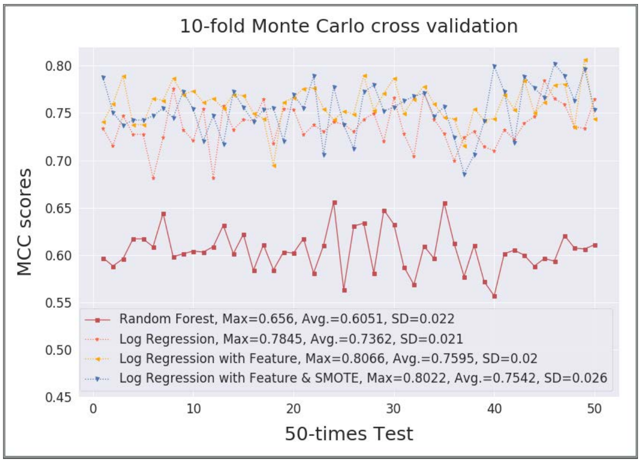


FIGURE 15. MCC accuracy comparison ablation experiments with 50 runs of 10-fold Monte Carlo cross-validation.

brought many benefits to the multi-category classification performance. Consequently, RF is selected to be the base classifier, because its standard deviation is smaller than other models in the experiments.

V. CONCLUSION

In this study, the researchers proposed Cascade RF–LR (with SMOTE) using the ABC algorithm to detect the multiclass probabilities of HCV incidence. This objective was achieved using a cascade two-stage method combining the RF and LR algorithms. The final results were a combination of the results obtained from the two models. The critical threshold value for separating Model 1 and Model 2 was obtained through optimized searching using the ABC algorithm. The proposed model was evaluated using various performance measurement indicators, including prediction accuracy, precision, recall, F1-score, and MCC. In addition, the proposed model was compared against the latest algorithms. The mean values obtained from 50 runs of 10-fold Monte Carlo cross-validation experiments were used as the retrieved values.

The results indicated that Cascade RF–LR (with SMOTE) using the ABC algorithm can be used to detect the multiclass probabilities of HCV, indicating that this model can be used to improve the effectiveness of relevant treatments. Despite the presence of imbalanced data in the clinical data of medical cases, the method in this combination was not considered in other combinations. Finally, improvements to prediction accuracy in situations involving insufficient IRs and samples will be explored in future research, such that the proposed method can be applied to the collection of complex data relating to rare diseases and medical treatments in clinical practice.

REFERENCES

[1] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: Management, analysis and future prospects," *J. Big Data*, vol. 6, no. 1, p. 54, Dec. 2019.

[2] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Amsterdam, The Netherlands: Elsevier, 2012.

[3] B. Hajarizadeh, J. Grebely, and G. J. Dore, "Epidemiology and natural history of HCV infection," *Nature Rev. Gastroenterol. Hepatol.*, vol. 10, no. 9, pp. 553–562, Sep. 2013.

[4] J. S. Bajaj, "Alcohol, liver disease and the gut microbiota," *Nature Rev. Gastroenterol. Hepatol.*, vol. 16, no. 4, pp. 235–246, Apr. 2019.

[5] J.-H. Wang, C.-H. Chen, C.-M. Chang, W.-C. Feng, C.-Y. Lee, and S.-N. Lu, "Hepatitis C virus core antigen is cost-effective in community-based screening of active hepatitis C infection in Taiwan," *J. Formosan Med. Assoc.*, vol. 119, no. 1, pp. 504–508, Jan. 2020.

[6] T. Boettler, P. N. Newsome, M. U. Mondelli, M. Maticic, E. Cordero, M. Cornberg, and T. Berg, "Care of patients with liver disease during the COVID-19 pandemic: EASL-ESCMID position paper," *JHEP Rep.*, vol. 2, no. 3, Jun. 2020, Art. no. 100113.

[7] J. H. Hoofnagle, "Hepatitis C: The clinical spectrum of disease," *Hepatology*, vol. 26, no. S3, pp. 15S–20S, Dec. 1997.

[8] A. M. Anter, S. Bhattacharyya, and Z. Zhang, "Multi-stage fuzzy swarm intelligence for automatic hepatic lesion segmentation from CT scans," *Appl. Soft Comput.*, vol. 96, Nov. 2020, Art. no. 106677.

[9] R. T. Ribeiro, R. T. Marinho, and J. M. Sanches, "Classification and staging of chronic liver disease from multimodal data," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 5, pp. 1336–1344, May 2013.

[10] G. Chandrasekaran, P. R. Karthikeyan, N. S. Kumar, and V. Kumarasamy, "Test scheduling of system-on-chip using dragonfly and ant lion optimization algorithms," *J. Intell. Fuzzy Syst.*, vol. 40, no. 3, pp. 4905–4917, Mar. 2021, doi: 10.3233/JIFS-201691.

[11] G. Chandrasekaran, S. Periyasamy, and P. R. Karthikeyan, "Test scheduling for system on chip using modified firefly and modified ABC algorithms," *Social Netw. Appl. Sci.*, vol. 1, no. 9, p. 1079, Sep. 2019, doi: 10.1007/S42452-019-1116-X.

[12] G. Chandrasekaran, V. Kumarasamy, and G. Chinraj, "Test scheduling of core based system-on-chip using modified ant colony optimization," *J. Européen Systèmes Automatisés*, vol. 52, no. 6, pp. 599–605, Dec. 2019, doi: 10.18280/JESA.520607.

[13] G. Chandrasekaran, G. Singaram, R. Duraisamy, A. S. Ghodake, and P. K. Ganesan, "Test scheduling and test time reduction for SoC by using enhanced firefly algorithm," *Revue d'Intell. Artificielle*, vol. 35, no. 3, pp. 265–271, Jun. 2021, doi: 10.18280/RIA.350310.

[14] J. H. Joloudari, H. Saadatfar, A. Dehzangi, and S. Shamshirband, "Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection," *Informat. Med. Unlocked*, vol. 17, Jan. 2019, Art. no. 100255.

[15] C.-C. Wu, W.-C. Yeh, W.-D. Hsu, M. M. Islam, P. A. A. Nguyen, T. N. Poly, Y.-C. Wang, H.-C. Yang, and Y.-C. J. Li, "Prediction of fatty liver disease using machine learning algorithms," *Comput. Methods Programs Biomed.*, vol. 170, pp. 23–29, Mar. 2019.

[16] J. Y. Nakayama, J. Ho, E. Cartwright, R. Simpson, and V. S. Hertzberg, "Predictors of progression through the cascade of care to a cure for hepatitis C patients using decision trees and random forests," *Comput. Biol. Med.*, vol. 134, Jul. 2021, Art. no. 104461.

[17] O. Kadioglu, M. Saeed, H. J. Greten, and T. Efferth, "Identification of novel compounds against three targets of SARS CoV-2 coronavirus by combined virtual screening and supervised machine learning," *Comput. Biol. Med.*, vol. 133, Jun. 2021, Art. no. 104359.

[18] R. Kumar, V. Kumar, and K. W. Lee, "A computational drug repurposing approach in identifying the cephalosporin antibiotic and anti-hepatitis C drug derivatives for COVID-19 treatment," *Comput. Biol. Med.*, vol. 130, Mar. 2021, Art. no. 104186.

[19] D. Chicco and G. Jurman, "An ensemble learning approach for enhanced classification of patients with hepatitis and cirrhosis," *IEEE Access*, vol. 9, pp. 24485–24498, 2021.

[20] A. Fabijanska and S. Grabowski, "Viral genome deep classifier," *IEEE Access*, vol. 7, pp. 81297–81307, 2019.

[21] H.-J. Chiu, T.-H.-S. Li, and P.-H. Kuo, "Breast cancer–detection system using PCA, multilayer perceptron, transfer learning, and support vector machine," *IEEE Access*, vol. 8, pp. 204309–204324, 2020.

[22] A. Basher, B. C. Kim, K. H. Lee, and H. Y. Jung, "Volumetric feature-based Alzheimer's disease diagnosis from sMRI data using a convolutional neural network and a deep neural network," *IEEE Access*, vol. 9, pp. 29870–29882, 2021, doi: 10.1109/ACCESS.2021.3059658.

[23] P. Khan, M. F. Kader, S. M. R. Islam, A. B. Rahman, M. S. Kamal, M. U. Toha, and K.-S. Kwak, "Machine learning and deep learning approaches for brain disease diagnosis: Principles and recent advances," *IEEE Access*, vol. 9, pp. 37622–37655, 2021, doi: 10.1109/ACCESS.2021.3062484.

[24] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2001, pp. 1–8.

[25] G. Heitz, S. Gould, A. Saxena, and D. Koller, "Cascaded classification models: Combining models for holistic scene understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 21, 2009, pp. 641–648.

[26] Z. Zhang and P. H. S. Torr, "Object proposal generation using two-stage cascade SVMs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 102–115, Jan. 2016.

[27] Z. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2195–2211, Sep. 2020.

[28] X. Liang, D. Li, M. Song, A. Madden, Y. Ding, and Y. Bu, "Predicting biomedical relationships using the knowledge and graph embedding cascade model," *PLoS ONE*, vol. 14, no. 6, Jun. 2019, Art. no. e0218264.

[29] N. Michielli, U. R. Acharya, and F. Molinari, "Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals," *Comput. Biol. Med.*, vol. 106, pp. 71–81, Mar. 2019.

[30] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[31] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[32] B.-H. Kung, P.-Y. Hu, C.-C. Huang, C.-C. Lee, C.-Y. Yao, and C.-H. Kuan, "An efficient ECG classification system using resource-saving architecture and random forest," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 6, pp. 1904–1914, Jun. 2021.

[33] S. Isci, D. S. Y. Kalender, F. Bayraktar, and A. Yaman, "Machine learning models for classification of Cushing's syndrome using retrospective data," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 8, pp. 3153–3162, Aug. 2021, doi: 10.1109/JBHI.2021.3054592.

[34] L. Luo, X. Yu, Z. Yong, C. Li, and Y. Gu, "Design comorbidity portfolios to improve treatment cost prediction of asthma using machine learning," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 6, pp. 2237–2247, Jun. 2021.

[35] M. Zhu, J. Xia, X. Jin, M. Yan, G. Cai, J. Yan, and G. Ning, "Class weights random forest algorithm for processing class imbalanced medical data," *IEEE Access*, vol. 6, pp. 4641–4652, 2018.

[36] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Reading, U.K.: Stanford Univ., 2019. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[37] D. Karaboga and B. Basturk, "On the performance of artificial bee colony (ABC) algorithm," *Appl. Soft Comput.*, vol. 8, pp. 687–697, Jan. 2008.

[38] D. Karaboga and B. Akay, "A comparative study of artificial bee colony algorithm," *Appl. Math. Comput.*, vol. 214, no. 1, pp. 108–132, Aug. 2009.

[39] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, Jan. 2002.

[40] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, pp. 113–141, Dec. 2013.

[41] G. Haixiang, Li Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.

[42] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2017.

[43] M. Ijaz, G. Alfian, M. Syafrudin, and J. Rhee, "Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest," *Appl. Sci.*, vol. 8, no. 8, p. 1325, Aug. 2018.

[44] Z.-H. Zhou and J. Feng, "Deep forest: Towards an alternative to deep neural networks," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 3553–3559.

[45] Z.-H. Zhou and J. Feng, "Deep forest," *Nat. Sci. Rev.*, vol. 6, no. 1, pp. 74–86, Jan. 2019.

[46] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[47] F. J. Rohlf and M. Corti, "Use of two-block partial least-squares to study covariation in shape," *Systematic Biol.*, vol. 49, no. 4, pp. 740–753, Dec. 2000.

[48] J. A. Wegelin, "A survey of partial least squares (PLS) methods, with emphasis on the two-block case," Univ. Washington, Seattle, WA, USA, Tech. Rep. 371, 2000.

[49] *HCV Data Set of the Machine Learning Repository in University of California, Irvine*. Accessed: May 20, 2022. [Online]. Available: https://archive-beta.ics.uci.edu/ml/datasets/HCV+data

[50] D. Dua and C. Graff. (2019). *UCI Machine Learning Repository*. University of California, School of Information and Computer Science. Irvine, CA, USA. [Online]. Available: http://archive.ics.uci.edu/ml

[51] G. Hoffmann, A. Bietenbeck, R. Lichtinghagen, and F. Klawonn, "Using machine learning techniques to generate laboratory diagnostic pathways—A case study," *J. Lab. Precis. Med.*, vol. 3, no. 6, p. 58, Jun. 2018.

[52] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[53] Z.-H. Zhou and X.-Y. Liu, "On multi-class cost-sensitive learning," *Comput. Intell.*, vol. 26, no. 3, pp. 232–257, Aug. 2010.

[54] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta-Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975.

[55] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–6, Dec. 2020.

[56] J. Gorodkin, "Comparing two K-category assignments by a K-category correlation coefficient," *Comput. Biol. Chem.*, vol. 28, nos. 5–6, pp. 367–374, Dec. 2004.

TZUU-HSENG S. LI (Member, IEEE) received the B.S. degree from the Tatung Institute of Technology, Taipei, Taiwan, in 1981, and the M.S. and Ph.D. degrees from the National Cheng Kung University (NCKU), Tainan, Taiwan, in 1985 and 1989, respectively, all in electrical engineering.

Since 1985, he has been with the Department of Electrical Engineering, NCKU, where he is currently a Distinguished Professor. From 1996 to 2009, he was a Researcher with the Engineering and Technology Promotion Center, National Science Council, Tainan. From 1999 to 2002, he was the Director of the Electrical Laboratories, NCKU. From 2009 to 2012, he was the Dean of the College of Electrical Engineering and Computer Science, National United University, Miaoli City, Taiwan. From 2009 to 2018, he was the Vice President of the Federation of International Robot-Soccer Association. He has been the Director of the Center for Intelligent Robotics and Automation, NCKU, since 2014. His current research interests include artificial and/or biological intelligence and applications, fuzzy system and control, home service robots, humanoid robots, mechatronics, 4WIS4WID vehicles, and singular perturbation methodology.

Dr. Li was elevated to CACS Fellow and a RST Fellow, in 2008 and 2018, respectively. He was elected as the President of the CACS, from 2008 to 2011, and the RST, from 2012 to 2015. He was a recipient of the Outstanding Automatic Control Award from the Chinese Automatic Control Society (CACS), Taiwan, in 2006, the Outstanding Research Award from the Ministry of Science and Technology, in 2016, and the Outstanding Robotics Engineering Award from the Robotics Society of Taiwan (RST), in 2017. He was a Technical Editor of the IEEE/ASME Transactions on Mechatronics and an Associate Editor of the *Asian Journal of Control*. He is currently the Editor-in-Chief of *International Journal of Robotics Research* and an Associate Editor of the *International Journal of Electrical Engineering*, the *International Journal of Fuzzy Systems*, and the IEEE Transactions on Cybernetics.

HUAN-JUNG CHIU received the B.S. degree from the Department of Mechanical and Marine Engineering, National Kaohsiung Marine Institute of Technology, Kaohsiung, Taiwan, in 2001, and the M.S. degree from the Department of Electrical Engineering, National Cheng Kung University (NCKU), Tainan, Taiwan, in 2012, where he is currently pursuing the Ph.D. degree. His current research interests include fuzzy control, intelligent systems, humanoid robot, image processing, machine learning, and robotic application.

PING-HUAN KUO (Associate Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, in 2008, 2010, and 2015, respectively. From 2017 to 2021, he was an Assistant Professor at the Department of Intelligent Robotics, National Pingtung University. Since 2022, he has been an Associate Professor at the Department of Mechanical Engineering, National Chung Cheng University. His major research interests include fuzzy control, intelligent algorithms, humanoid robot, image processing, robotic application, big data analysis, machine learning, and deep learning applications.